Parallel-QC users' manual

Version 1.0

Introduction

Quality control (QC) is the critical first step for processing raw next-generation sequencing (NGS) data, in which sequencing artifacts, including low-quality reads and contaminating reads, would significantly affect and mislead the results of downstream analysis. Here we report Parallel-QC, a fast computational engine specifically designed for general NGS data QC. Parallel-QC can complete sequencing-quality assessment accuracy and efficiency. Possible contaminating species could also be identified without any prior information. And the whole processing of Parallel-QC is quite fast since it is optimized based on parallel computation.

Download

The latest version can be downloaded at:
http://www.computationalbioenergy.org/parallel-qc.html

Package Dependency

Linux GCC 4.1 or higher
POSIX Thread Library

Install

Extract the package:

```
tar –xzvf parallel-qc-1.0.tar.gz
```

Compile the source code:

```
cd parallel-qc-1.0
make
make rand-sel
```

Tools in toolkit

parallel-qc

The main tool for quality control.

rand-sel

An random reads extraction tool.

Usage

parallel-qc

The **parallel-qc** accepts pair-ended or un-pair-ended sequences in FASTA and FASTQ format.

parallel-qc [Options] Value

[Options]:

-I Input file name(s) **[Required]**
Input file must be in FASTQ or FASTA format, supporting 1 (single-ended sequences, or pair-ended sequences in single file) or 2 (pair-ended sequences in separated files) names.

-Q quality file name(s)
Must have the same IDs as the input sequences. Supporting 1 (single-ended sequences, or pair-ended sequences in single file) or 2 (pair-ended sequences in separated files, in the same order as the input file names) names. Available only when the input is in FASTA format.

-o Output file directory **[Required]**

-f Output format. Q or F
Q for FASTQ and F for FASTA. Default is FASTQ

-p T or F
If the input is pair-end reads. Default is T.

-k T or F
If keep the pairs for output. Default is T.

-d T or F
If drop the duplicated reads. Default is T

-b begin(int) end(int)
Base trim begin and end. Keep 0 0 to turn off. Default is turn off.

-q quality(int) ratio(float, 0~1), default in Sanger format. Add '-P' if in Phred format.
Trim quality value and threshold ratio that minimum percentage of bases must have the trim quality value. Keep 0 0 to turn off. Default is turn off

-m Tag sequences file, must be in FASTA format. Maximum sequence number is 100. Default is turn off. The recommended tag sequences are located at Parallel-QC_directory/Default_tag_sequence/primer-combined.fa.

-g GC min_ratio(float, 0~1) max_ratio(flaot, 0~1).
GC proportion trim, must between min_ratio ~ max_ratio. Default is turn off

-t thread number. Default thread number is 1.
-h Print help.

rand-sel

The **rand-sel** accepts pair-ended or un-pair-ended sequences in FASTA and FASTQ format.

rand-sel [Options] Value

[Options]:

-I Input file name(s) **[Required]**
Input file must be in FASTQ or FASTA format, Supporting 1 (single-ended sequences, or pair-ended sequences in single file) or 2 (pair-ended sequences in separated files) names.

-Q quality file name(s)
Must have the same IDs as the input sequences. Supporting 1 (single-ended sequences, or pair-ended sequences in single file) or 2 (pair-ended sequences in separated files, in the same order as the input file names) names. Available only when the input is in FASTA format.

-o Output file name prefix **[Required]**
-f Output format Q or F
Q for FASTQ and F for FASTA. Default is FASTQ
-p T or F
If the input is pair-end reads. Default is T.
-P select_ratio(float, 0-1)
Random selection proportion.
-h Print help.

Results

parallel-qc

For parallel-qc, all analysis results will be in the directory assigned by parameter '**–o'**.

If the output reads are not kept into pair-ended, the suffix of '-1' and '-2' will be removed for each pair of output files. In the output directory, files are

read-1.fa & read-2.fa: The output sequence file after quality control analysis.

trim-qual-1.fa & trim-qual-2.fa: Trimmed reads by quality trimming step.

trim-primer-1.fa & trim-primer-2.fa: Trimmed reads by tag sequence trimming.

trim-gc-1.fa & trim-gc-2.fa: Trimmed reads by GC proportion trimming.

trim-dup-1.fa & trim-dup-2.fa: Trimmed reads by duplication trimming.

analysis_report.txt: The overall information of the quality control analysis.

rand-sel

For rand-sel, the randomly selected reads files are named by the parameter '**-o**' and suffix '-1' and '-2' for pair-ended output. Suffix names are also removed for un-pair-ended output.

Notice

1. The output path will be cleared initially, and please make sure parallel-qc has the write permission of the output path.
2. Make sure the input is in fasta or fastaq format, and select the correct option.
3. Parameter for quality file '–Q' is available only when the input file(s) are in FASTA format.
4. Please assign '-p F' if the input file is not pair-ended.
5. For tag sequence trimming, maximum tag sequence number is 100.
6. The recommended tag-sequences are located at Parallel-QC_directory/Default_tag_sequence/primer-combined.fa.

Contact

Any problems please feel free to contact

Dr. Kang Ning
ningkang@qibebt.ac.cn

Xiaoquan Su
suxq@qibebt.ac.cn

Dr. Qian zhou
zhouqian@qibebt.ac.cn